

# **EXHIBIT 14**

# LLM Data Deep Dive

**NOTE: SENSITIVE DATA — DO NOT SHARE BROADLY**

Discussion: 7/5/23

[Pre-read doc](#)

## Background, definitions & intuition

## Our unit of measurement is a *token*

- A piece of text, usually char < **token** < word
- Breaks down documents into machine readable chunks, e.g:

untokenized: "Breaks down documents into machine readable  
chunks"

tokenized: ['\_Bre', 'aks', '\_down', '\_documents', '\_into',  
'\_machine', '\_readable', '\_ch', 'unks']

## Data scale intuitions: pretraining

We use  $O(\text{Trillions})$  of unlabeled tokens for pretraining

To contextualize 1T tokens of text =

- ~12M books
- ~double all the podcast episodes ever made
- ~40B social media posts or ~160 days of FB group posts

## Data scale intuitions: pretraining vs tuning

Reasonable tuning can happen with just 1K labeled examples (e.g. prompt-response pairs). *Millions of times less data than pretraining.*

Our fine-tuning uses ~1M labeled examples. *Thousands of times less data than pretraining.*

Mix of:

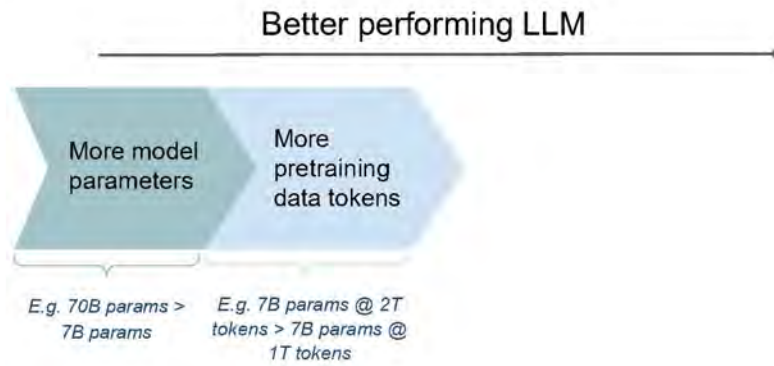
- Supervised fine-tuning: O(10k - 100k) prompt<>response pairs, O(10M-100M) tokens.
- Reward model data: O(100k - 1M) prompt<>response<>ratings, O(100M-1B) tokens.

Some new behaviors can be learned well with O(10s) of SFT examples. *Hundreds of millions times less data than pretraining.*

## Data scale intuitions: basic scale laws

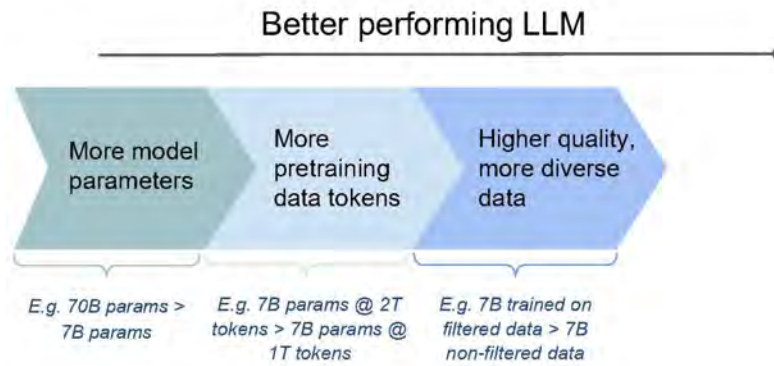


## Data scale intuitions: basic scale laws

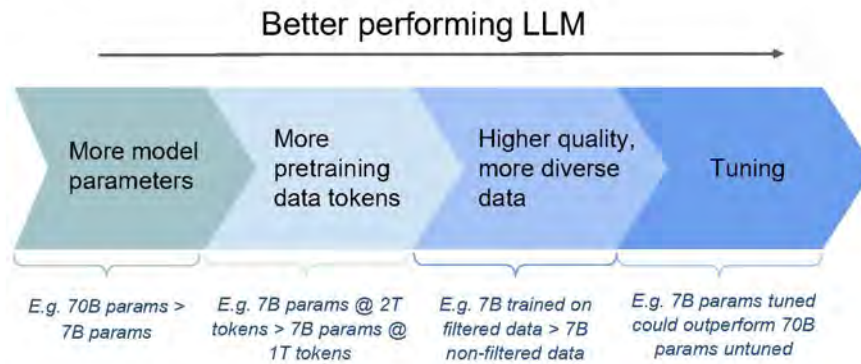




## Data scale intuitions: basic scale laws



## Data scale intuitions: basic scale laws



Best models have it all...

## Data Timelines

Getting data ready for a big run is a usually a multi-(3+)month process

**Acquire/Approve**  
*Weeks to months*

**Download /preprocess  
(OCR, parse) data**  
*Days to weeks*

**Measure + analyze**  
*Days to weeks*

**Move data**  
*Hours to days*

**Apply Mitigations**  
*Hours to days*

**Process Data + Run Experiments:** iterative process, dedupe/clean, rank; often involves implementing techniques from literature or novel techniques.  
*Days to months*

**Cross-corpus experiments**  
*Days to months*

Current state: data, models and compute

## Content of our current models

LLaMa-Anise-70B

**1.4T tokens (w/ epochs)**

**1.3T tokens (raw)**

- Common Crawl: 82.3%
- Wikipedia: 4.5%
- Code: 4.5%
- Books: 4.5%
- Arxiv: 2.4%
- Stack Exchange: 1.8%

## Content of our current models

**LLaMa-Anise-70B**  
**1.4T tokens (w/ epochs)**  
**1.3T tokens (raw)**

- Common Crawl: 82.3%
- Wikipedia: 4.5%
- Code: 4.5%
- Books: 4.5%
- Arxiv: 2.4%
- Stack Exchange: 1.8%

**LLaMa-Cinnamon-70B**  
**2T tokens (w/ epochs)**  
**1.8T tokens (raw)**

- Common Crawl: 82.2%
  - Wikipedia: 4.8%
  - Code: 5.6%
  - Books: 4.4%
  - Arxiv: 1.6%
  - Stack Exchange: 1.2%
- [includes data mitigations]*

+0.5T new tokens of more recent  
common crawl. Otherwise proportions  
similar to de-risk.

## Content of our current models

### LLaMa-Anise-70B 1.4T tokens (w/ epochs) 1.3T tokens (raw)

- Common Crawl: 82.3%
- Wikipedia: 4.5%
- Code: 4.5%
- Books: 4.5%
- Arxiv: 2.4%
- Stack Exchange: 1.8%

### LLaMa-Cinnamon-70B 2T tokens (w/ epochs) 1.8T tokens (raw)

- Common Crawl: 82.2%
  - Wikipedia: 4.8%
  - Code: 5.6%
  - Books: 4.4%
  - Arxiv: 1.6%
  - Stack Exchange: 1.2%
- [includes data mitigations]*

### LLaMa-Dill-150B 2.3T tokens (w/ epochs) 2T tokens (raw)

- Common Crawl: 74.3%
  - Code: 11.6%
  - Wikipedia: 4.3%
  - Books: 3.6%
  - Arxiv: 2.8%
  - Stack Exchange: 2.2%
  - Math: 1.1%
- [includes data mitigations]*

+0.5T new tokens of more recent  
common crawl. Otherwise proportions  
similar to de-risk.

Add math, more code data

## How does Llama stack up?

*More data than others at same params, but less data than larger parameter size models*

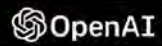
Falcon (7B, 40B) <b>1-1.5T tokens</b>	LLaMa-Anise (7B, 13B, 30B, 65B): <b>1-1.4T tokens</b>	GPT-4 (8x220B? 1T?): <b>15T Tokens??</b> text + image
PaLM-2 (340B): <b>3.6T tokens</b>	LLaMa-Cinnamon (7B, 13B, 33B, 70B): <b>2T tokens</b>	
	LLaMa-Dill (150B): <b>2.3T tokens</b>	



## Data sources: Competitive landscape



- Ranked web crawl
- Google book project
- Google docs?
- YouTube for audio/  
multimodal/ transcribed  
speech
- Google photos



- Ranked web crawl
- Microsoft office?
- Photos?



Meta

- Instagram
- FB
- Whatsapp

## We are significantly lagging GPT-4 and PaLM-2

<b>Benchmark</b> <i>Higher numbers are better.</i>	<b>ChatGPT API</b> (est. 175B)	<b>GPT-4</b> (8x220B? 1T?)	<b>PALM (540B)</b>	<b>PALM-2-L</b> (est. 340B)	<b>Llama (70B)</b>
Knowledge: MMLU	70.0	<b>86.4</b>	69.3	78.3	68.9
Knowledge: TriviaQA	N/A	N/A	81.4	86.1	71.5
Knowledge: Natural Questions	N/A	N/A	29.3	37.5	33.0
Mathematical Reasoning: GSM8K	57.1	<b>92.0</b>	56.5	80.7	56.8
Coding: HumanEval	48.1	<b>67.0</b>	26.2	N/A	29.9

Closing the gap requires training more powerful base models with more data.

## Compute needs per params<>tokens

Model-Params	Tokens needed	Compute Needed**
Cinnamon-70B	2T	2K A100s for 6 weeks
Dill-150B	2.3T	4K A100s for 2 months
5X Dill-250B?	8-10T	8K A100s for 5 months
GPT-4 equivalent-340B?	12-15T?*	16K A100s for 8 months
LLaMA-Next-340B?	20T?*	16K H100s for 6 months

\* Probably need multimodal data as well to compete.

\*\* Not counting compute needs for ablations/evals; add ~30%

## Compute needs per params<>tokens

Model	Params	Tokens needed	Compute Needed**
C D 5	<b>Reality check:</b> GPT-4 finished training ~October 2022. 16K A100s is all of RSC, not likely to happen ASAP? If our big run starts in Nov (optimistic) on H100s, we'll "only" be ~18 months behind GPT-4 training.		16K A100s for 6 weeks
			16K A100s for 2 months
			16K A100s for 5 months
	GPT-4 equivalent-340B?	12-15T?*	16K A100s for 8 months
	LLaMA-Next-340B?	20T?*	16K H100s for 6 months

\* Probably need multimodal data as well to compete.

\*\* Not counting compute needs for ablations/evals; add ~30%

Moar data

## Short-term data goal: enough data to train LLama-Next

**Old:** Llama-Dill  
Size: **2.3T (2T raw)**

**Current:** Lots of new raw text data,  
mix of quality

Size: **~83T tokens**  
>40X increase in size

**Post-processing  
expectation:** High  
quality, mixed-  
lingual  
Size: **10-15T  
tokens**

**Goal:** High quality,  
mixed-lingual,  
multimodal corpus  
Size: **20T tokens?**

## Corpus content: data sources





## Corpus content: quality

### Example low quality data

Welcome to the Calthness.Org Forums.  
 - ## Specials Threads / Posts Last Post  
 - ##### Forum Actions: ##### Forum Statistics:  
 - Threads: 4  
 - Posts: 8  
 ##### Last Post: by 01-May-09, 16:51  
 - A forum for you to test out the features of the board, or discuss "how do I ...?" issues.  
 ##### Forum Actions: ##### Forum Statistics:  
 - Threads: 303  
 - Posts: 1,366  
 ##### Last Post: by 29-Aug-20, 21:16  
 - ## General Threads / Posts Last Post  
 - ##### Forum Actions: ##### Forum Statistics:  
 - Threads: 34,812  
 - Posts: 511,065  
 ##### Last Post: by 14-Apr-21, 13:23  
 - ##### Forum Actions: ##### Forum Statistics:  
 - Threads: 636  
 - Posts: 3,138  
 ##### Last Post: by 21-Feb-21, 21:43  
 - ##### Forum Actions: ##### Forum Statistics:

<http://forum.calthness.org/index.php?s=f994e42d16a4ecb13756051e2790159>

### Example high quality data

#### # FDA 101: Dietary Supplements

Dietary supplements can help improve health but can also have risks. Get the facts on supplements and how the FDA regulates them to help keep you safe.

Multivitamins, vitamin D, echinacea, and fish oil are among the many dietary supplements lining store shelves or available online. Perhaps you already take a supplement or are thinking about using one. Dietary supplements can be beneficial to your health, but they can also involve health risks. So, it's important that you talk with a health care professional to help you decide if a supplement is right for you.

Read on to learn what dietary supplements are (and are not), what role the U.S. Food and Drug Administration has in regulating them, and how to make sure you and your family use supplements safely.

#### ## What Are Dietary Supplements?

Dietary supplements are intended to add to or supplement the diet and are different from conventional food. Generally, to the extent a product is intended to treat,

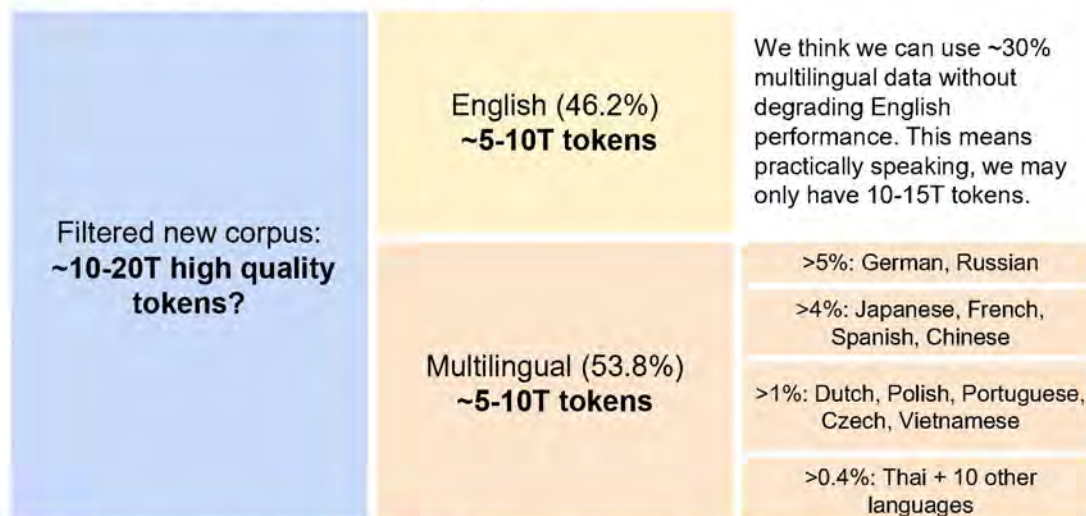
<https://www.fda.gov/consumers/consumer-update/fda-101-dietary-supplements>

Higher quality/more diverse tokens lead to the same training results faster + with less compute. We estimate ~10-15T will be "high quality", which we find by:

- Deduping, processing
- Ranking on quality signals
- Mixing in different categories, languages, context-lengths

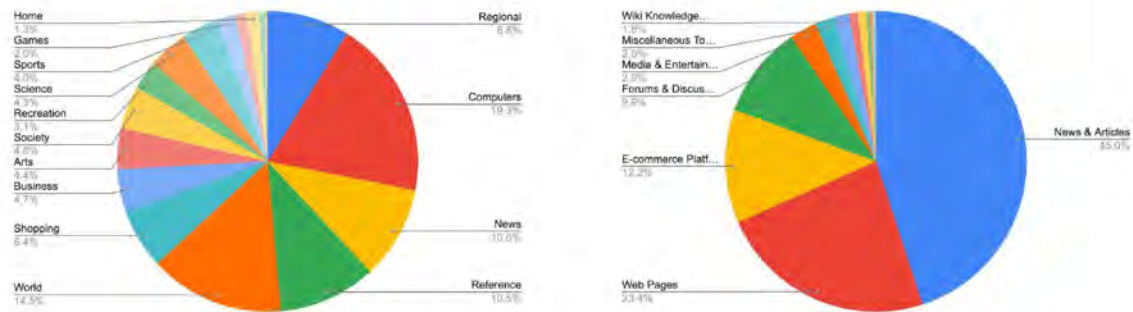


## Corpus content: languages



Note: these are estimates from [common crawl stats](#); we haven't analyzed all our own data fully yet.

## Corpus content: topics and types



Percentage of tokens in CommonCrawl (English)

# Redacted

Future work, challenges and investment areas

## New data sources

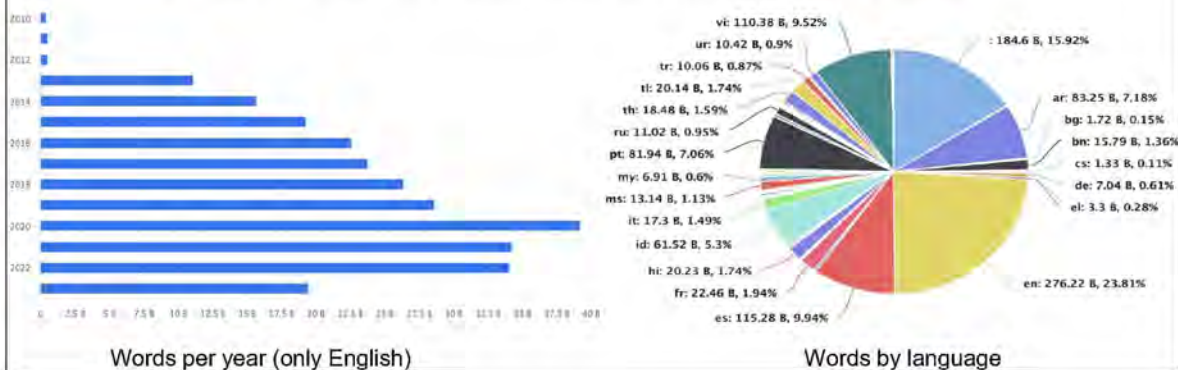
Audio, Video Speech -> Text:  
**1-2T tokens**

Prod data (biz messages, posts, comments): **1-5T tokens**

Multimodal: 5T? Image tokens, 100sB text tokens?

## Internal data: 1-5T(?) high quality tokens

- US FB messenger contains 500M *words* per day- averaging 180B words/year
  - Initial fine-tuning experiments on small sets running now (500K threads with > 5 turns)
- FB data ingestion in progress: Public *posts* contain 1.2T words.
  - Working to implement PII mitigations. Comment ingestion currently in progress.



## FB MSGR fine-tuning

**Prompt:** *My mom is mad at me for not doing my homework, what should I do?*

**new\_base\_80pct\_response:** If your mom is mad at you for not doing your homework, it's important to take responsibility for your actions and make things right. Here are some steps you can take:

1. Acknowledge your mistake: Admit that you made a mistake by not doing your homework and apologize for it.
2. Explain the reason: If there was a valid reason for not doing your homework, explain it to your mom. However, make sure you are not making excuses or blaming others.
3. Offer a solution: Suggest a plan to catch up on your homework and prevent it from happening again in the future.
4. Show willingness to learn: Express your willingness to learn from your mistake and make an effort to improve your study habits.
5. Ask for help: If you need help with your homework, ask your mom or a tutor for assistance.
6. Follow through: Make sure you follow through on your plan and complete your homework to your best ability.

Remember, it's important to be honest and take responsibility for your actions. By doing so, you can earn back your mom's trust and respect.

**new\_base\_p2b\_1500\_response:** I think you should try to explain the situation to her and let her know how you're feeling. It might help to show her that you're taking your schoolwork seriously, even if you're not doing it in the traditional way. Maybe you can work out a compromise with her?

Early fine-tuning experiment by Aasish

## Challenges

### Data acquisition speed

- Takes months to acquire (new) data sources: legal slowdowns, lack of people, lack of scaleable infra
  - Examples: FoA data acquisition has been taking months; new efforts like speech/podcasts will likewise take months.
- No clear multimodal strategy

### Development speed & process

- Lack of repeatable & fast data processes; code is often one-off / AWS codebase
- Data processing tooling (dedup etc.) still immature

### Data insights

- Lack of data domain analysis; lack of tooling and people
  - Impacts data downselection and data acquisition goal setting



## Investments needed:

Acquisition [\[Research/eng/product leads, PXFN\]](#)

- Focus on speech to text data; reconsider more code
- Watch laws and see how we need to adjust
- More reviews for internal data

## Investments needed:

### Acquisition [\[Research/eng/product leads, PXFN\]](#)

- Focus on speech to text data; reconsider more code
- Watch laws and see how we need to adjust
- More reviews for internal data

### Science + Processing [\[Research, Foundations\]](#)

- Experiments on deduplication, data mix, data quality, data x RAI, data scale laws. [\[Research\]](#)
- Many of these require implementation work first [\[Research, Foundations\]](#)

## Investments needed:

### Acquisition [\[Research/eng/product leads, PXFN\]](#)

- Focus on speech to text data; reconsider more code
- Watch laws and see how we need to adjust
- More reviews for internal data

### Science + Processing [\[Research, Foundations\]](#)

- Experiments on deduplication, data mix, data quality, data x RAI, data scale laws. [\[Research\]](#)
- Many of these require implementation work first [\[Research, Foundations\]](#)

### Tools/Infra [\[Foundations, Infra, DE\]](#)

- Data processing libraries that accelerate (processing time, researcher time) processing techniques, allow for better analyses

## Investments needed:

### Acquisition [\[Research/eng/product leads, PXFN\]](#)

- Focus on speech to text data; reconsider more code
- Watch laws and see how we need to adjust
- More reviews for internal data

### Science + Processing [\[Research, Foundations\]](#)

- Experiments on deduplication, data mix, data quality, data x RAI, data scale laws. [\[Research\]](#)
- Many of these require implementation work first [\[Research, Foundations\]](#)

### Tools/Infra [\[Foundations, Infra, DE\]](#)

- Data processing libraries that accelerate (processing time, researcher time) processing techniques, allow for better analyses

### Analysis [\[DE, DS\]](#)

- Analytics to help us rank/filter training data, acquire new data, improve pretraining safety

## Investments needed:

### Acquisition [\[Research/eng/product leads, PXFN\]](#)

- Focus on speech to text data; reconsider more code
- Watch laws and see how we need to adjust
- More reviews for internal data

+ Connecting acquisition, infra, processing across modalities [\[research, platform, infra, DE\]](#)

### Science + Processing [\[Research, Foundations\]](#)

- Experiments on deduplication, data mix, data quality, data x RAI, data scale laws. [\[Research\]](#)
- Many of these require implementation work first [\[Research, Foundations\]](#)

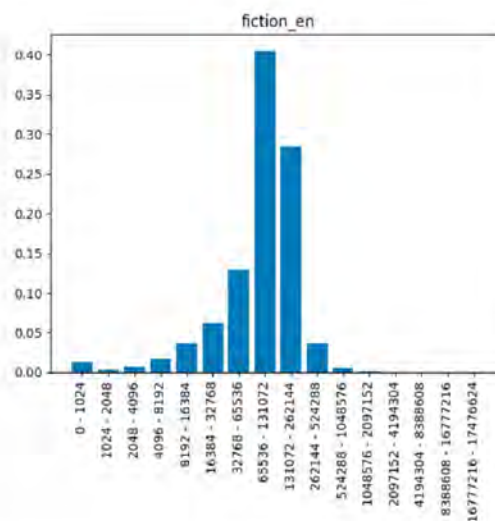
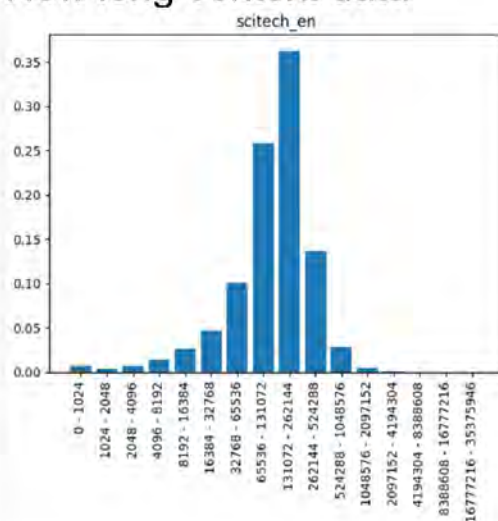
### Tools/Infra [\[Foundations, Infra, DE\]](#)

- Data processing libraries that accelerate (processing time, researcher time) processing techniques, allow for better analyses

### Analysis [\[DE, DS\]](#)

- Analytics to help us rank/filter training data, acquire new data, improve pretraining safety

## New long context data



## Responsible AI at the data layer

- Generally, do light scrubbing: high risk PII/IP sites, don't scrub toxic/biased content
  - Idea is that models that see more bad content in pretraining generalize to better safety in tuning
- In the future:
  - Explore tradeoffs/balances more: especially for adult content, memorization, PII
  - Experiment with control / source tokens and other ways to label training data for downstream tuning + explainability